MOLE-BLAST a new tool to search and classify multiple sequences

Greg Boratyn, Christiam Camacho, Scott Federhen, Yuri Merezhuk, Tom Madden, Conrad Schoch, and Irena Zaretskaya

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

Abstract:

MOLE-BLAST is a new tool to classify multiple query sequences and discover their relationship to each other. This tool provides a taxonomic context for the queries. First, MOLE-BLAST groups the query sequences by locus. Second, it performs a BLAST database search to identify each query's nearest neighbors. Third, it computes a multiple alignment for each locus, including query sequences and their nearest neighbors. Finally, MOLE-BLAST presents the result of its analysis as phylogenetic tree for each locus.

MOLE-BLAST is available at:
http://blast.ncbi.nlm.nih.gov/blast/moleblast/moleblast.cgi

MOLE-BLAST is useful for:

- Taxonomists: to find sequence neighbors and their taxonomic context
- Ecologists
- Pathologists
- Scientists submitting sequences to NCBI: to verify that sequences have correct taxonomic annotation

A user can:

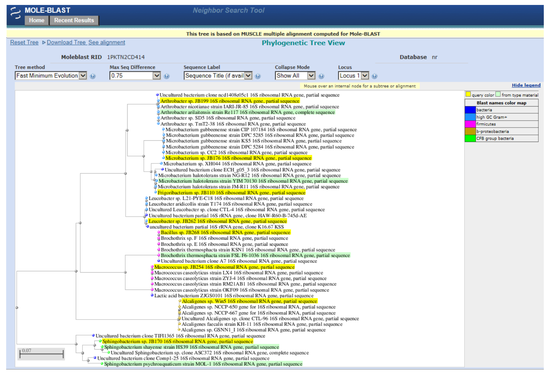
- Quickly find neighbor sequences
- Assess sequence membership in taxonomic groups
- Find taxonomic context of one's sequences
- Separate a large set of sequences into different genes or loci
- Visualize relationships to sequences from type reference specimens

Example:

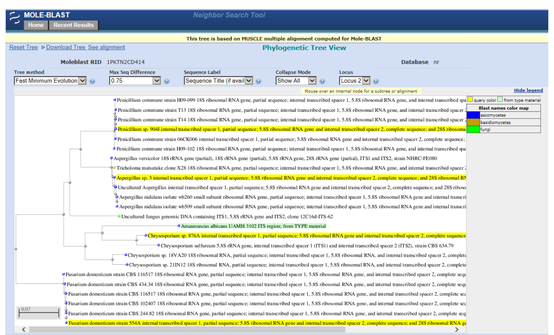
We submitted to MOLE-BLAST a subset of 16S and ITS sequences from [1] that presents a study of microbial communities collected from the surfaces of naturally aged cheese. The results are the phylogenetic trees below for 16S and ITS sequences. The user can easily visualize the relationship of the bacterial and fungal query sequences to one another and to BLAST [2] search results.

User's query sequences are highlighted in yellow. Additionally, database sequences from type reference material are highlighted in green.

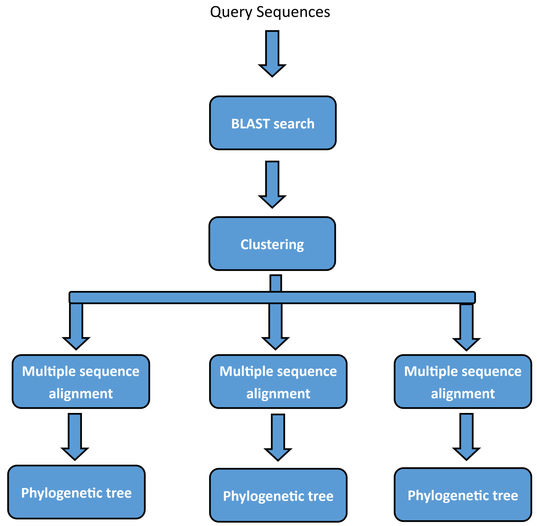
Result for locus 1:



Result for locus 2:



MOLE-BLAST work flow



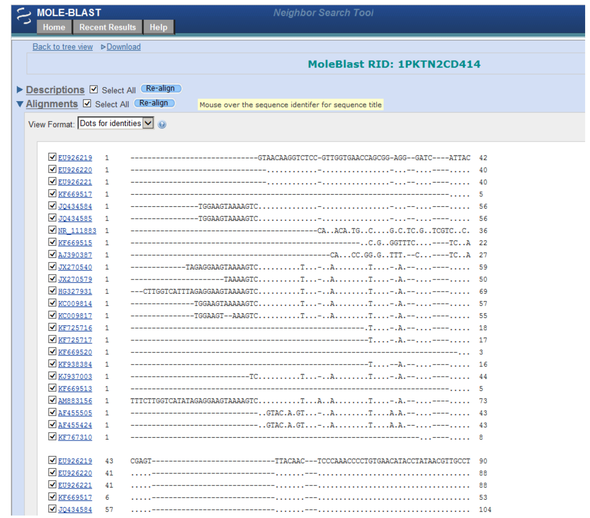
BLAST search: A user can search queries against the NR, RefSeq RNA, or 16S Ribosomal RNA databases. The search can also be restricted with an Entrez query. Top five BLAST search results for each query sequence are included for further processing.

Clustering: The input sequences are aligned to one another with BLAST and groups of sequences that all match to one another (cliques) form a cluster .

Multiple sequence alignment: Query sequences that belong to a cluster along with their top BLAST search results are submitted to Muscle [3] for multiple alignment. Database sequences shorter than the length of the longest query (plus 20 bases) are included without modification; longer sequences are trimmed to the extent covered by BLAST alignment.

Phylogenetic tree: A phylogenetic tree is computed for each locus multiple sequence alignment using Neighbor Joining [4] or Fast Minimum Evolution [5].

The user can also examine the multiple sequence alignment used to compute the phylogenetic tree.



Future features:

- Manual correction of multiple alignment extent
- Easy selection and downloading of sequences
- Prioritization of database sequences
- Highlighting sequences from verified material
- MOLE-BLAST for proteins

References:

[1] Wolfe BE, Button JE, Santarelli M and Dutton RJ (2014) Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity, *Cell* 158:422-433

[2] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389-402

[3] Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792-1997

[4] Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406-425

[5] Desper R and Gascuel O (2004) Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol Biol Evol.* 21:587-298